Published in final edited form as:
Proteins. 2014 October ; 82(10): 2565–2573. doi:10.1002/prot.24620.

Direct prediction of profiles of sequences compatible to a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles

Zhixiu Li[1], **Yuedong Yang**[1,2,3], **Eshel Faraggi**[1,2,4,5,6], **Jian Zhan**[1,2,3], and **Yaoqi Zhou**[1,2,3,*]

[1]School of Informatics and Computing, Indiana University-Purdue University, Indianapolis, Indiana 46202, USA

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

[3]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr. Southport, QLD 4222, Australia

[4]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[5]Battelle Center for Mathematical Medicine, Nationwide Children's Hospital, Columbus, Ohio 43215, USA

[6]Research and Information Systems, LLC Carmel, IN 46032, USA

Abstract

Locating sequences compatible to a protein structural fold is the well-known inverse protein-folding problem. While significant progress has been made, the success rate of protein design remains low. As a result, a library of designed sequences or profile of sequences is currently employed for guiding experimental screening or directed evolution. Sequence profiles can be computationally predicted by iterative mutations of a random sequence to produce energy-optimized sequences, or by combining sequences of structurally similar fragments in a template library. The latter approach is computationally more efficient but yields less accurate profiles than the former because of lacking tertiary structural information. Here we present a method called SPIN that predicts Sequence Profiles by Integrated Neural network based on fragment-derived sequence profiles and structure-derived energy profiles. SPIN improves over the fragment-derived profile by 6.7% (from 23.6% to 30.3%) in sequence identity between predicted and wild-type sequences. The method also reduces the number of residues in low complex regions by 15.7% and has a significant better balance of hydrophilic and hydrophobic residues at protein surfaces. The accuracy of sequence profiles obtained is comparable to those generated from the protein design program RosettaDesign 3.5. This highly efficient method for predicting sequence profiles from structures will be useful as a single-body scoring term for improving scoring functions used in protein design and fold recognition. It also complements protein design programs in guiding

Corresponding author: Yaoqi Zhou, Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr. Southport, QLD 4222, Australia, yaoqi.zhou@griffith.edu.au.

Conflict of Interest: none declared.

experimental design of the sequence library for screening and directed evolution of designed sequences. The SPIN server is available at http://sparks-lab.org.

Keywords

Protein design; knowledge-based energy function; neural network; sequence profiles; inverse protein folding problem

1 INTRODUCTION

Designing a protein sequence that would fold into a given structure is the well-known inverse-protein folding problem. Solving this problem will not only improve our fundamental understanding of the interactions responsible for protein folding and structure prediction but also advance our capability of designing novel proteins with existing function improved or with completely new functionality.

Significant progress in protein design has been made in recent years with a number of designed sequences successfully validated experimentally in terms of their structures and their functions [1–11]. These designs typically start from random protein sequences and iteratively optimize an energy score via mutations until the scoring function reaches a minimum. However, existing scoring functions for protein design are not yet accurate enough to produce high success rates [12–16]. In fact, designed sequences usually do not contain wild-type sequences as a part of the solution [17,18]. Low success rate of single sequence design has led to current effort in employing multiple computationally predicted sequences (or sequence profiles) to build a sequence library for large-scale experimental screening of desirable properties [19–23] or for directed evolution [14,24]. Sequence or sequence profiles obtained from protein design programs require solving a *NP*-hard combinatorial optimization problem[25]. Thus, it is time consuming to produce sequence profiles based on multiple runs.

In addition to above energy-based methods, sequence profiles can also be predicted by employing local fragment structures [26]. In this approach, fragment structures from a target structure are compared to the fragment structures from a template library of known protein structures. Sequences of those template fragment structures with high structural similarity to target fragments are obtained to produce the sequence profile for the entire target structure by a sliding-widow approach. Sequence profiles generated from fragment structures and/or from protein design programs have been found useful for enhancing the ability of recognizing structural similarity in the absence of sequence similarity (fold recognition) by matching a sequence profile of a query not only with the sequence profile of a template sequence but also the sequence profile predicted from a template structure [26–28]. More recently, sequence profiles derived from fragment structures were employed as a single-body energy term for improving the energy function of protein design [17]. Predicting sequence profiles by fragments only needs to perform pairwise structural alignment between short fragments and, thus, is computationally much more efficient than solving the combinatorial optimization problem required by an energy-based design. However, sequence profiles derived from short fragments are dominated by local structural

information. That is, they are only useful for capturing the interactions responsible for local structure formation, but do not account for non-local interactions (interactions between structural but not sequence neighbors) that are responsible for the stability of tertiary structure. As a result, fragment-derived profiles are not as useful as the profiles derived from energy optimization for using in experimental screening or directed evolution.

In this paper, we test the idea of using neural network (NN) to improve fragment-derived sequence profiles by incorporating a mean-field like non-local interaction. We found that an energy-based nonlocal feature makes a significant improvement in the quality of sequence profiles over that from fragment structural alignment in terms of sequence identity to wild-type sequences, fraction of hydrophilic residues, recovery rate of wild-type residue types, precision of predicted amino-acid residue types, distribution of amino-acid residue types, and fraction of low complexity regions. The quality of predicted sequence profiles is comparable to the profiles generated from the protein design program RosettaDesign 3.5[29] based on several measures. This NN-derived profile is complementary to existing energy-based techniques for identifying sequences that are compatible with a desired structural fold. It should be also useful as a single-body term for improving the fold-recognition scoring function or protein-design energy function as fragment-based profiles did[17,26–28].

2 METHODS

2.1 Data sets

To perform training and test and avoid over-training, we need three datasets: structural templates, a dataset for training the neural network, and a dataset for independent test. For the template library, we started from a non-redundant protein set with resolution better than 2.0 Å, pair-wise sequence identity of less than 30% from the PISCES server [30] downloaded on October 17, 2008. This set contains 4803 protein chains that were further reduced to 2528 chains after removing chains with missing residues or backbone atoms. We further cleaned the dataset by removing proteins (1) complexed with DNA or RNA, (2) whose sequence contain un-recognized residue types; and (3) whose secondary structures were not defined by DSSP [31]. This leads to a total of 2282 protein chains that are employed as our templates for fragment structures (TL2282).

For training and test sets, we started from the new non-redundant protein set with resolution better than 3.0 Å, pair-wise sequence identity of less than 30% from the PISCES server[30] on April 28, 2013. This set contains 10460 protein chains. We cleaned the dataset using the same criteria above and removed all chains with >30% sequence identity to the proteins in the template library (TL2282). This leads to a dataset of 2032 proteins. We randomly selected 500 proteins for independent test (TS500) and utilized the remaining proteins for training and ten-fold cross validation (TR1532).

From TS500, we randomly selected 50 small proteins with sequence length between 60–200 and fraction of surface residue between 0.5–0.8 (TS50). This small dataset is used to compare the sequence profiles generated from our neural-network approach with those generated from RosettaDesign, one of the most widely used programs for protein design[29]. A small dataset is used because it is computationally intensive to produce sequence profiles

by designing 1000 sequences utilizing RosettaDesign. These 50 proteins (PDB ID plus chain ID) are 1eteA, 1v7mV, 1y1lA, 3pivA, 1or4A, 2i39A, 4gcnA, 1bvyF, 3on9A, 3vjzA, 3nbkA, 3l4rA, 3gwiA, 4dkcA, 3so6A, 3lqcA, 3gknA, 3nngA, 2j49A, 3fhkA, 2va0A, 3hklA, 2xr6A, 3ii2A, 2cayA, 3t5gB, 3ieyB, 3aqgA, 3q4oA, 2qdlA, 3ejfA, 3gfsA, 1ahsA, 2fvvA, 2a2lA, 3nzmA, 3e8mA, 3k7pA, 3ny7A, 2gu3A, 1pdoA, 1h4aX, 1dx5I, 1i8nA, 2cviA, 3a4rA, 1lpbA, 1mr1C, 2xcjA, and 2xdgA.

To remove all information from wild-type sequences in their structures, amino-acid residue types in the PDB structural files of all datasets were labeled as ALA (alanine). All native positions of C_β atoms were removed and replaced by the positions of pseudo C_β atoms based on standard 1.54 Å for the $C_\alpha - C_\beta$ bond length, 109.538° for the $N - C_\alpha - C_\beta$ bond angle and 109.468° for the $C - N - C_\alpha - C_\beta$ dihedral angle. All protein structures are not energy minimized prior to removal of original side chains to avoid possible "memory" of side chains by the energy function used in minimization. The latter could lead to artificially high sequence identity to wild type sequences.

2.2 Neural network

We employed the same neural-network method developed for sequence-based continuous-value prediction of backbone torsion angles and residue solvent accessibility [32–34]. It contains a two-hidden-layer neural network. Each of the two hidden layers contains 51 hidden neurons and one bias. We employed a bipolar activation function given by $f(x) = \tanh(\alpha x)$, with $\alpha = 0.2$. Back propagation with momentum was applied to optimize the weights. The learning rate and momentum were set to 0.001 and 0.4, respectively.

2.3 Input features

Local features—There are two types of local features. The first one is backbone torsion angles (φ and ψ) at a given sequence position. The second one is the fragment-derived sequence profile. The method for obtaining the fragment-derived sequence profile was described in [17]. Briefly, 5-residue fragments (from i to i+4; i=1, 2, …, L-4) in a target structure of sequence length L are structurally compared to all fragments in the same length located in the structural template library (TL2282). The sequences of most structurally similar fragments (in RMSD) are utilized to calculate probability of a residue type at each sequence position (sequence profile). For each sequence position, this profile has a dimension of twenty for 20 residue types.

Energy-based non-local features—We introduced an energy-based non-local feature as follows. For a given sequence position, we built the full side-chain based on the rotamers of each amino-acid residue type, one rotamer at a time while assuming that the residue type at all other positions is alanine. The total interaction energies of the residue of 20 residue types in all rotameric states with all other alanine residues are calculated separately. We record only the lowest total energy in all rotameric states of each residue type at a given sequence position plus the energies of six most frequent rotamers (or less if a residue type has less than six rotamers). The total number of features is 114 (=7×13+4×4+3×1+2×2) because four residue types have only three rotamers, Proline has two, and Glycine and Alanine have one conformation]. Here, the bbdep02 rotamer library [35] and a knowledge-

based energy function based on the distance-scaled finite-ideal gas reference state (DFIRE) [36,37] were employed.

Sliding window and normalization of input values—In addition to the features from the current position (i), we also include the features from two sequence neighbors (i−1 and i +1). That is, a window size of 3 is employed. We utilized this window size because a larger window size did not improve our prediction. The values of all input features were linearly transformed to [−1, 1]. The total number of input features is 136×3 (136=2+20+114).

2.4 Output

The output layer contains 20 nodes with each node representing one amino-acid residue type. In other words, the neural network outputs 20 probability values for 20 amino acid residue types for each sequence position. We trained the neural network to make two types of predictions. The first one is to predict wild-type sequences where each sequence is represented by a 20×L matrix. That is, each sequence position has a 20-dimension vector for 20 amino-acid residue types. The value is 1 if a particular residue type is located at the sequence position and −1 for all other dimensions. The second one is to predict position-specific substitution matrix (PSSM) generated by PSIBLAST[38]. This prediction takes into account the fact that more than one sequence can have the same structure. In this case, a 20×L matrix generated from PSIBLAST[38] is used as the target for training and prediction.

2.5 Ten-fold cross validation and independent test

To examine the accuracy of prediction, we performed 10-fold cross validation on TR1532. The dataset is randomly divided into 10 equal parts. Nine were used for training and the remaining was for testing. This process was repeated 10 times, once for each of the 10 parts. To prevent over-training, a random over-fit protection set with 5% of the training set is excluded from training and is used as a small test set for determining the stop criterion for neural-network weight optimization. We did 10 fold cross-validations for five times with different random seeds. The consensus of predicted amino-acid types of 5 independent runs is employed to calculate the sequence identity to wild-type sequences. For independent test, TR1532 was employed for training and TS500 was for test only.

2.6 Performance evaluation

The objective function in the neural network is to minimize the difference between predicted and actual values (20-dimension 1 and −1 vector or PSSM). The performance, on the other hand, is assessed by several different measures. One is the sequence identity between predicted sequence and the wild-type sequence, which is equal to the number of correctly predicted residue types divided by the total number of residues. We also calculated precision and recovery rate of each residue type where precision is the fraction of correctly predicted residues for a given residue type in the number of predicted residues of that type. Recovery rate is the fraction of correctly predicted residues of a given residue type in the number of wild-type residues of that type.

Another measure of performance is mean square error. In order to calculate the mean square error between PSSM and a predicted profile, the predicted profile (fragment and single-

sequence NN-based approaches, or RosettaDesign) was transformed to a pseudo PSSM by $\log(P_{ij})$, where P_{ij} is the probability for given residue type i in position j. Both pseudo PSSM and PSSM are normalized from 0 to 1. The mean square error is obtained by calculating the difference between PSSM and the best linear fit of the pseudo PSSM to the PSSM.

2.7 RosettaDesign

RosettaDesign 3.5 was downloaded from https://www.rosettacommons.org/software/. Proteins are designed based on a fixed backbone structure with the command "fixbb.linuxgccrelease -s example.pdb -resfile example.resfile -ex1 -ex2 -nstruct 100 -database ROSETTA_DATABASE -linmem_ig 10 -extrachi_cutoff 0 -ignore_unrecognized_res -no_opth false -skip_set_reasonable_fold_tree -no_his_his_paire -score:weights score12prime.wts". 1000 sequences were designed by optimizing all residues simultaneously for each protein in order to obtain a sequence profile. All positions are set as ALLAA in example.resfile. All structures are not minimized prior to optimization for design.

3 RESULTS

3.1 Sequence prediction

One way to measure the accuracy of design is to estimate the sequence identity between designed sequence and the original wild-type sequence. The fragment-based approach yields an average sequence identity of 23.6% for TR1532, which is consistent with 24% obtained by using other databases [17]. For the neural-network (NN) based approach, we can predict the "best" sequence based on the residue type that has the highest predicted value at each sequence position. We found that neural-network based prediction made a 7.1% improvement from 23.6% to 30.7% over the fragment-based approach. We can also evaluate the improvement based on top 2 predicted residue types. A correct prediction is made if one of the top 2 predictions matches to the wild-type sequence. The improvement is 8% from 36.3% by the fragment-based approach to 44.3% by the neural-network-based approach. For the independent test (TS500), the improvement is essentially identical at 7.1% (23.6% to 30.7%) for top 1 and 7.7% (36.1% to 43.8%) for top 2 matching, respectively.

To examine the relative importance of different features, we evaluated different combinations of three features employed here. Because we would like to compare against the fragment-based approach, we utilized the structure fragment profile as a base feature and added torsion angles or the energy-based profile for comparison. We found that adding the energy-based profile improves the sequence identity to wild-type sequences by 6% while adding the dihedral angles adds 1.4% only. In addition, using the energy-based profile alone can yield an average sequence identity of 26% to wild type sequences which is 2% higher than the fragment-based profile. These results highlight the importance of nonlocal interaction energy function in neural-network learning.

Figure 1 compares average sequence identities as a function of protein lengths (number of amino acid residues). The bins for protein lengths are [0–100), [100–200), and etc. The last bin contains all proteins with greater than 700 amino acid residues for TR1532 and greater

than 600 residues for TS500. The figure reveals a consistent improvement of the neural-network based prediction over the fragment-based prediction for different sizes of proteins. Moreover, the result from the independent test is nearly indistinguishable from the ten-fold cross validation, highlighting the robustness of our training method.

Because it is more difficult to design regions exposed to water, it is useful to examine how sequence identity will change for proteins with different fractions of surface residues. A residue is defined as on surface if its solvent accessible surface is greater than or equal to 20% of its reference value. All proteins were divided into 12 bins according to fractions of surface residues ([0.35–0.4), [0.4, 0.45), [0.45, 0.5), [0.5, 0.55), [0.55, 0.6), [0.6, 0.65), [0.65, 0.7), [0.7, 0.75),[0.75–0.8), [0.8–0.85), [0.85–0.9) ,[0.9,1]). Because the dataset TS500 does not have enough data to form the bin [0.9,1], we combined those proteins to the bin [0.85–0. 9). We started from a fraction of 0.35 because all proteins contain at least 35% surface residues. Figure 2 displays the average sequence identity as a function of the fraction of surface residues in a protein. Consistent with other methods [17,18], sequence identities between predicted and actual sequences are lower for proteins with higher fraction of surface residues. Again, there is a consistent improvement of 2–10% by the neural-network-based method over the fragment-based method regardless the value of the fraction of the surface residues. We further observed the consistency between the ten-fold cross validation and the independent test.

We calculated the recovery rate and precision for each residue type. As shown in Figure 3A, the NN-based approach improves over the fragment-based approach in 15 out of 20 residue types for both precision and recovery rate. We noted that glycine (G) and proline (P) are the most accurately predicted residue types because of their unique backbone conformations. Recovery rates for R (Arg), H (His), Q (Glu), C (Cys), M (Met), and W (Trp) for both approaches are very low. This behavior is likely due to low occurrence of residue types such as W, M, C, and H in wild-type sequences. Figure 3B compares the occurrence of 20 amino acid residue types in wild-type sequences with those in predicted sequences. We calculated the Kullback–Leibler divergence of residue distribution between NN approach and wild-type and that between fragment-based approach and wild-type sequences. The former is 0.18 and the latter is 0.31. That is, the NN approach yields a distribution much closer to that of wild-type sequences than the fragment-based approach except for residue E (Glu) where the NN approach over-predicts it. We found that the NN approach over-predicts E because it often mis-predicts R and Q as E. 27.8% Q residues were predicted as E, 13.6% as K and 11% as L. 20.8% of R residues were predicted as E, 15.3% as K and 12.2% as L. The confusion between R and Q (both under-predict) with E and K (both over-predict) are likely due to the fact that all of them are hydrophilic residues with relatively long side-chains.

Table I further examines sequence identity in different secondary structure and in surface regions (only independent test results shown as they are essentially same as ten-fold cross validation). Interestingly, coil regions in protein backbones have the highest identity (30% by fragment and 35% by neural network), compared to 26% in helical or 25% in sheet regions. This is largely because of high occurrence of Gly and Pro in coil regions. These two residue types were most accurately predicted because of their unique backbone conformations. The most significant improvement of the NN approach over the fragment-

based approach is in the core region (10.5% increase in sequence identity). Table I also shows the fraction of hydrophilic residues. It is clear that the NN approach has a significantly better balance of hydrophilic-hydrophobic residues on the surface of proteins in particular (34% by the fragment-based approach, 60% by the NN approach and 65% in wild-type sequences). However, there is no improvement in the core of proteins which have 10% less hydrophilic residues in predicted sequences than in wild-type sequences. Here hydrophilic residues refer to D, E, H, K, N, Q, R, S, T, and Y.

Low complexity region (e.g. multiple repeats of same residue type such as VVV) is often associated with intrinsically disordered regions of proteins. We have employed the program SEG[39] to locate low complexity regions in predicted sequences. As Table I shows, the fraction of residues in low complexity regions is as high as 50.8% per protein by the fragment-based approach for the test set TS500. The NN approach cuts it to 34.5%, although it is still significantly higher than 3% in wild-type sequences.

3.2 PSSM Prediction

So far, we have trained our NN to predict a single sequence despite the fact that there are more than one sequence that could be fitted for a single structure. Thus, it is of interest to know if training a NN to predict sequence profile directly, rather than a single sequence, would lead to an improved result. To do this, we use the Position Specific Substitute Matrix (PSSM) generated from PSIBLAST[38] for training and testing the NN approach. The PSSM is normalized to −1 to 1. We define a PSSM consensus sequence based on the most frequent residue from PSSM at each sequence position.

Table II compares sequence identities between consensus sequences from PSSM and predicted consensus sequences by the fragment-based approach, the NN trained by single sequence and the NN trained by PSSM. Interestingly, the NN trained by PSSM is similar to the NN trained by a single sequence when judged by the sequence identity to the PSSM consensus sequence (26.6% versus 26.1% for TR1532 and 26.3% versus 26.7% for TS500 for top 1). Improvement on the mean square error (MSE) is greater because the NN trained by PSSM was directly optimized for MSE. The difference in conserved regions between NN (single sequence) and NN (PSSM) is also small. For example, the sequence identity to the consensus sequence in the conserved regions (PSSM≥7) is 31.8% by single-sequence trained NN (single sequence) and 32.4% by PSSM-trained NN.

We compared the fractions of hydrophilic residues in PSSM consensus sequences and in wild-type sequences and found that they are quite similar (28.4% in PSSM consensus sequence versus 27.9% in wild-type sequence in protein core and 61.3% in PSSM consensus sequence versus 64.3% in wild-type sequences in protein surface). However, the PSSM trained NN predicts significantly more hydrophilic residues (5%) on protein surface and 3% more in protein core than the single-sequence trained NN. It is unclear why using PSSM for training neural networks would significantly increase the number of hydrophilic residues on the surface of proteins.

3.3 Comparison to profiles generated by RosettaDesign

We compared to RosettaDesign[29] for 50 proteins due to costly computational requirement by using RosettaDesign for producing sequence profiles. As shown in Table III, RosettaDesign deviates more from wild-type PSSM than NN-based approaches do. Its sequence identity to wild-type sequence (based on the average sequence identity from 1000 designed sequences) is similar to the NN-based approach. Interestingly, RosettaDesign employs significantly more hydrophilic residues in core than wild-type sequences while fragment-based and NN-based approaches consistently under-predict hydrophilic residues in the core. RosettaDesign, however, has similar number of residues in low complexity regions as wild-type sequences, as it was optimized for.

4 Discussion

In this paper, we employed neural networks for predicting sequences associated with a given protein structure. We found that a local fragment-derived sequence profile can be significantly improved by integrating with an energy-based nonlocal feature through neural networks. Together with backbone torsion angles, the neural-network based method SPIN makes 7% improvement over fragment-derived sequence profiles in sequence identity to wild-type sequences. The accuracy of sequence profiles from SPIN is comparable to RosettaDesign in term of sequence identity to wild-type sequences and sequence variation. The MSE between predicted and actual PSSM given by single-sequence trained SPIN is 0.198, compared to 0.223 by RosettaDesign for a dataset of 50 proteins. SPIN and RosettaDesign also yield similar sequence identities to wild-type sequences (~30%).

The average 30% sequence identity for 50 proteins achieved by RosettaDesign is significantly lower than 37.0% reported by Leaver-Fay et al[40] despite the same scoring function and procedures were employed. A close examination found that this discrepancy is caused by structural relaxation prior to sequence design. Structural relaxation of crystal structures by RosettaDesign prior to design inevitably introduces the bias toward wild-type sequences and lead to a higher sequence identity. We found that for the 50 proteins, relaxation prior to design yielded an average sequence identity of 35.6%. Here, we reported the results from RosettaDesign without pre-relaxation to be consistent with the structures employed for SPIN.

SPIN can be considered as a mean-field like approach. This is because nonlocal interaction energy is calculated by assuming that all neighboring residues except the residue of interest are alanine. We used alanine because it is the smallest amino acid residue except glycine. Using a residue with a small side chain is necessary to avoid steric clashes. We do not utilize glycine because lacking a side chain makes it different from most residue types by allowing a much more flexible backbone conformation. Moreover, alanine has only one conformation. Thus, there is no need for optimizing its rotameric state. In addition, alanine is the second most widely employed amino acid residues in proteins (8.1%, only 1% behind 9.5% for leucine). The abundance level in protein structures is important for minimizing the error caused by approximating all other positions as alanine. It should be mentioned that using alanine for the energy-based nonlocal profile brings over-predicted alanine (19%) by fragment-based profile to a population (7%) similar to the actual population (8%).

The comparable accuracy between SPIN and RosettaDesign suggests that there is room for further improving an energy-based approach. In fact, thirty percent sequence identity to wild-type sequence reached by this neural-network method and the difficulty to improve much beyond 30% for protein design by energy optimization [17,18] suggests a common bottleneck facing protein design. This 30% sequence identity is in a so-called twilight zone[41] where two protein sequences may or may not have the same structure [17]. That is, going beyond 30% is necessary to significantly improve the success rate of protein design. Typical energy functions for protein design contain, at minimum, single-body profiles and two-body pairwise interaction terms. In contrast, SPIN relied on single body energetic terms only. Thus, SPIN raises the bar for protein design programs that are based on more sophisticated energetic terms. On the other hand, the results of SPIN can be effectively employed as a single-body energy term to improve an energy function for design. In our previous work, we found that incorporation of the fragment-derived profile into the RosettaDesign energy function [42] can increase the sequence identity by 4–8% [17]. Using this newly improved profile (7% higher sequence identity over the fragment-based approach) as an energy term may further improve the ability of recovering wild-type sequences.

Another potential application of this structure-derived profile is fold recognition. Several studies have found that sequence profiles from protein design significantly improve the ability of recognizing structural similarity in the absence of sequence similarity [26–28]. This is particularly important for recognizing new structure folds that do not have wild-type sequence information but are generated from multiple loop permutations [43]. Application to fold recognition is feasible because SPIN is computationally efficient. It takes only 343 processor seconds to predict one sequence profile from structures, compared to 833×1000 processor seconds by RosettaDesign for predicting 1000 sequences by Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60ghz.

There is a recent trend to overcome low success rate of design by using a library of protein sequences designed by a design program. The library is then utilized for large-scale experimental screening of desirable properties [19–23] or for directed evolution [14,24]. SPIN provides a complementary approach to protein design programs for building a library of sequences that are compatible to a given structure with similar accuracy at a much lower computational cost.

One way to further improve SPIN is to improve its energy-based features. The nonlocal energy profile was obtained by employing a DFIRE-based statistical energy function. We employed this energy function because it has been found useful in protein structure and binding prediction and other applications [44]. Other coarse-grained statistical potentials (backbone only) [45] can also be employed here. Obviously, DFIRE or any other statistical energy functions were not optimized for this purpose. One might expect that our method can be further improved if a knowledge-based potential is optimized for single-residue-type recovery when the rest proteins are approximated as occupied by alanine residues.

One surprising finding is that using PSSM to train neural networks does not lead to any visible improvement over the single-sequence based training. Essentially the same sequence identity to PSSM consensus sequences is observed despite that the single-sequence method

was not trained for predicting PSSM at all. In fact, we found that the top two amino acid residue types predicted by single-sequence-trained NN are essentially the same as the top two amino acid residue types by the PSSM-trained NN (87.5% in agreement). This suggests that a neural network is capable of capturing the profile encoded in a given protein structure regardless if it was trained or not trained by a profile. In other words, the structure of a protein has a dominated effect on the evolution of sequences.

Acknowledgments

Funding: This work was supported by National Institutes of Health R01 GM085003, a Project Development Team within the ICTSI NIH/NCRR Grant Number TR000006, and National Health and Medical Council of Australia (Grant number 1059775) to Y.Z and by Windows Azure for Research Award to YY. We also gratefully acknowledge the support of the Griffith University eresearch Services Team and the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

References

1. Dahiyat BI, Mayo SL. De novo protein design: Fully automated sequence selection. Science. 1997; 278(5335):82–87. [PubMed: 9311930]
2. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science. 1998; 282(5393):1462–1467. [PubMed: 9822371]
3. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF. From coiled coils to small globular proteins: Design of a native-like three-helix bundle. Protein science : a publication of the Protein Society. 1998; 7(6):1404–1414. [PubMed: 9655345]
4. Walsh ST, Cheng H, Bryson JW, Roder H, DeGrado WF. Solution structure and dynamics of a de novo designed three-helix bundle protein. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96(10):5486–5491. [PubMed: 10318910]
5. Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL. Full-sequence computational design and solution structure of a thermostable protein variant. J Mol Biol. 2007; 372(1):1–6. [PubMed: 17628593]
6. Bender GM, Lehmann A, Zou H, Cheng H, Fry HC, Engel D, Therien MJ, Blasie JK, Roder H, Saven JG, DeGrado WF. De novo design of a single-chain diphenylporphyrin metalloprotein. J Am Chem Soc. 2007; 129(35):10732–10740. [PubMed: 17691729]
7. Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded beta-sheet protein. Science. 1998; 281(5374):253–256. [PubMed: 9657719]
8. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. Accurate computer-based design of a new backbone conformation in the second turn of protein l. J Mol Biol. 2002; 315(3):471–477. [PubMed: 11786026]
9. Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, Francois JM, Mayo SL, Martial JA. De novo backbone and sequence design of an idealized alpha/beta-barrel protein: Evidence of stable tertiary structure. J Mol Biol. 2003; 325(1):163–174. [PubMed: 12473459]
10. Dobson N, Dantas G, Baker D, Varani G. High-resolution structural validation of the computational redesign of human u1a protein. Structure. 2006; 14(5):847–856. [PubMed: 16698546]
11. Dantas G, Corrent C, Reichow SL, Havranek JJ, Eletr ZM, Isern NG, Kuhlman B, Varani G, Merritt EA, Baker D. High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J Mol Biol. 2007; 366(4):1209–1221. [PubMed: 17196978]
12. Suarez M, Jaramillo A. Challenges in the computational design of proteins. Journal of the Royal Society Interface. 2009; 6:S477–S491.
13. Lippow SM, Tidor B. Progress in computational protein design. Current opinion in biotechnology. 2007; 18(4):305–311. [PubMed: 17644370]

14. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science. 2011; 332(6031):816–821. [PubMed: 21566186]
15. Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C, Friedman B, Hanf KJM, Jarpe M, Liparoto SF, Li Y, Lugovskoy A, Miller S, Rushe M, Sherman W, Simon K, Van Vlijmen H. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. Protein Sci. 2006; 15(5):949–960. [PubMed: 16597831]
16. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, Chan C, Chung HS, Eivazi A, Yoder SC, Vielmetter J, Carmichael DF, Hayes RJ, Dahiyat BI. Engineered antibody fc variants with enhanced effector function. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(11):4005–4010. [PubMed: 16537476]
17. Dai L, Yang Y, Kim HR, Zhou Y. Improving computational protein design by using structure-derived sequence profile. Proteins. 2010; 78(10):2338–2348. [PubMed: 20544969]
18. Li ZX, Yang YD, Zhan J, Dai L, Zhou YQ. Energy functions in de novo protein design: Current challenges and future prospects. Annu Rev Biophys. 2013; 42:315–335. [PubMed: 23451890]
19. Hayes RJ, Bentzien J, Ary ML, Hwang MY, Jacinto JM, Vielmetter J, Kundu A, Dahiyat BI. Combining computational and experimental screening for rapid optimization of protein properties. Proc Natl Acad Sci U S A. 2002; 99(25):15926–15931. [PubMed: 12446841]
20. Treynor TP, Vizcarra CL, Nedelcu D, Mayo SL. Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. Proc Natl Acad Sci U S A. 2007; 104(1):48–53. [PubMed: 17179210]
21. Guntas G, Purbeck C, Kuhlman B. Engineering a protein-protein interface using a computationally designed library. Proc Natl Acad Sci U S A. 2010; 107(45):19296–19301. [PubMed: 20974935]
22. Allen BD, Nisthal A, Mayo SL. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc Natl Acad Sci U S A. 2010; 107(46):19838–19843. [PubMed: 21045132]
23. Chen TS, Palacios H, Keating AE. Structure-based redesign of the binding specificity of anti-apoptotic bcl-x(l). J Mol Biol. 2013; 425(1):171–185. [PubMed: 23154169]
24. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature. 2008; 453(7192):190–195. [PubMed: 18354394]
25. Pierce NA, Winfree E. Protein design is np-hard. Protein Eng. 2002; 15(10):779–782. [PubMed: 12468711]
26. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 2005; 58(2):321–328. [PubMed: 15523666]
27. Busch MSA, Mignon D, Simonson T. Computational protein design as a tool for fold recognition. Proteins. 2009; 77(1):139–158. [PubMed: 19408297]
28. Larson SM, Garg A, Desjarlais JR, Pande VS. Increased detection of structural templates using alignments of designed sequences. Proteins-Structure Function and Genetics. 2003; 51(3):390–396.
29. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97(24): 13383–13388.
30. Wang G, Dunbrack RL Jr. Pisces: A protein sequence culling server. Bioinformatics. 2003; 19(12): 1589–1591. [PubMed: 12912846]
31. Kabsch W, Sander C. Dictionary of protein structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
32. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. Spine x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of computational chemistry. 2012; 33(3):259–267. [PubMed: 22045506]

33. Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins. 2009; 74(4):847–856. [PubMed: 18704931]
34. Faraggi E, Yang YD, Zhang SS, Zhou Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure. 2009; 17(11):1515–1527. [PubMed: 19913486]
35. Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997; 6(8):1661–1681. [PubMed: 9260279]
36. Zhou HY, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002; 11(11): 2714–2726. [PubMed: 12381853]
37. Yang YD, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci. 2008; 17(7):1212–1219. [PubMed: 18469178]
38. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: A new generation of protein database search programs. Nucleic Acids Research. 1997; 25(17):3389–3402. [PubMed: 9254694]
39. Wootton JC, Federhen S. Statistics of local complexity in amino-acid-sequences and sequence databases. Computers & chemistry. 1993; 17(2):149–163.
40. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song YF, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B. Scientific benchmarks for guiding macromolecular energy function improvement. Methods in Protein Design. 2013; 523:109–143.
41. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12(2):85–94. [PubMed: 10195279]
42. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol. 2003; 332(2):449–460. [PubMed: 12948494]
43. Dai L, Zhou Y. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. J Mol Biol. 2011; 408:585–595. [PubMed: 21376059]
44. Zhou Y, Zhou HY, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys. 2006; 46(2):165–174. [PubMed: 17012757]
45. Tozzini V. Coarse-grained models for proteins. Curr Opin Struc Biol. 2005; 15(2):144–150.

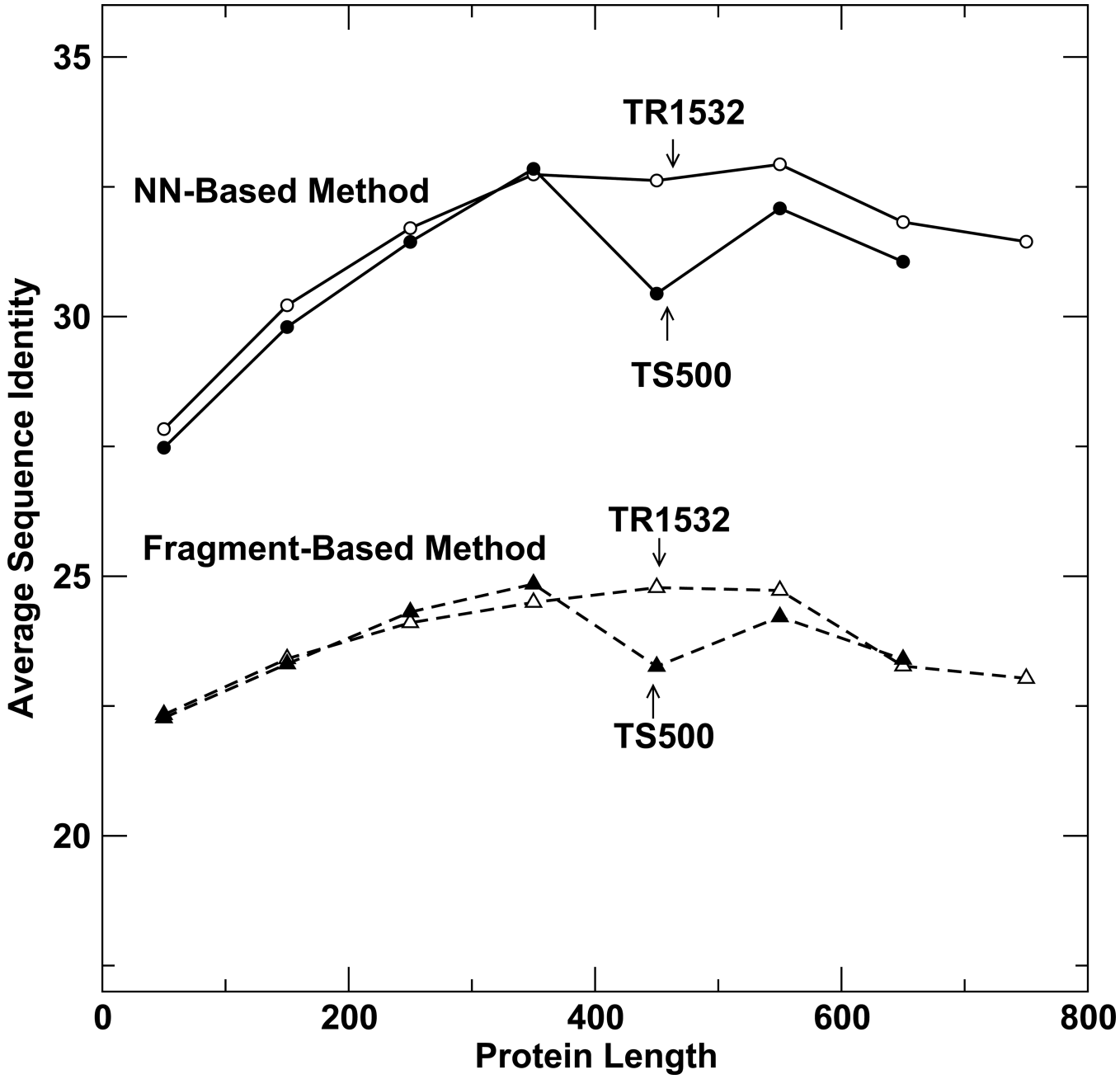

Figure 1.

Average sequence identity between predicted and wild-type sequences as a function of protein length (ten-fold cross validation on TR1532, open symbols and independent test on TS500, filled symbols) by the fragment-based (dashed lines) and neural-network based approaches (solid lines).

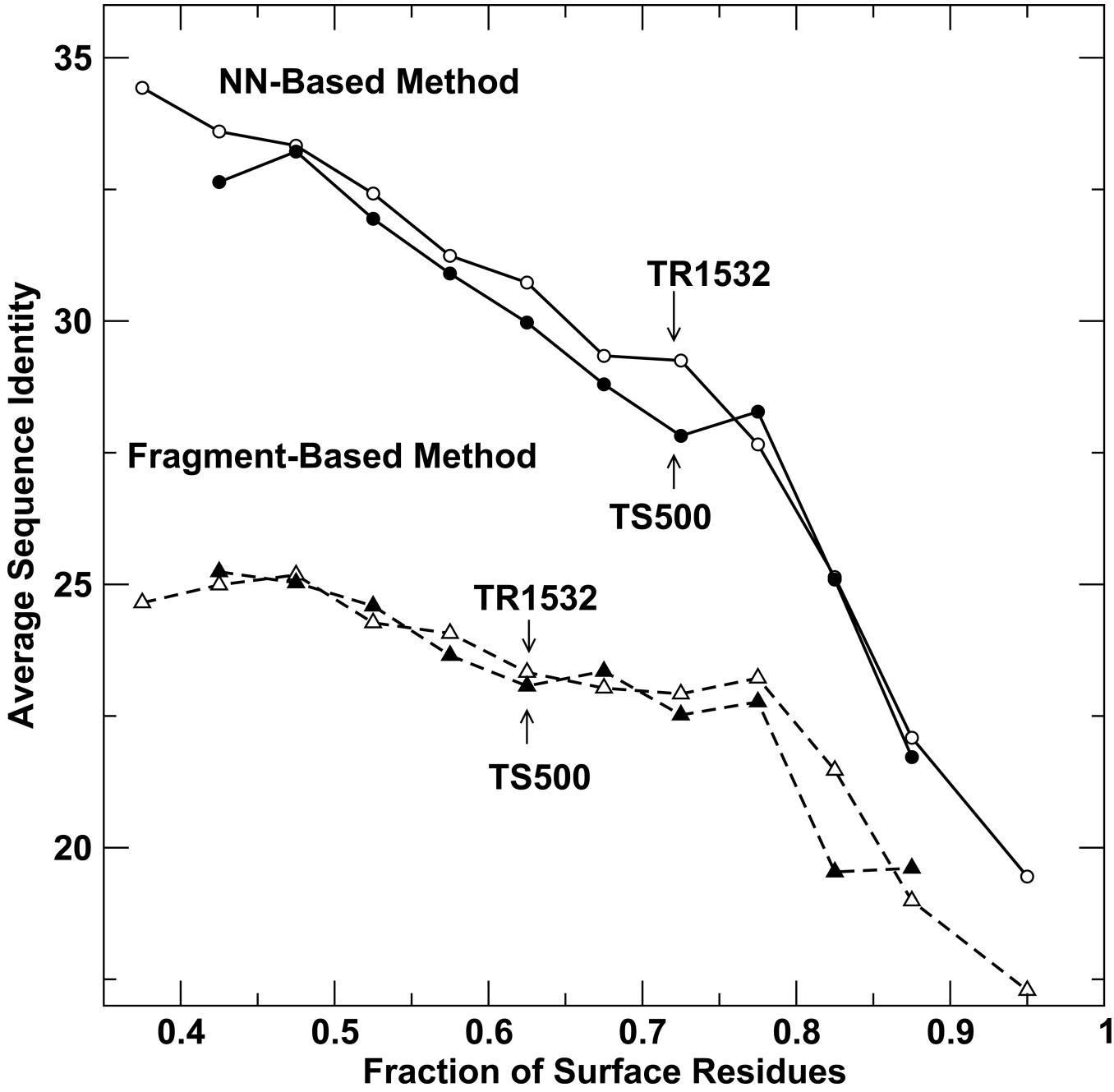

Figure 2.

Average sequence identity between predicted and wild-type sequences as a function of the fraction of surface residues (ten-fold cross validation on TR1532, open symbols and independent test on TS500, filled symbols) by the fragment-based (dashed lines) and the neural-network (NN) based approaches (solid lines).

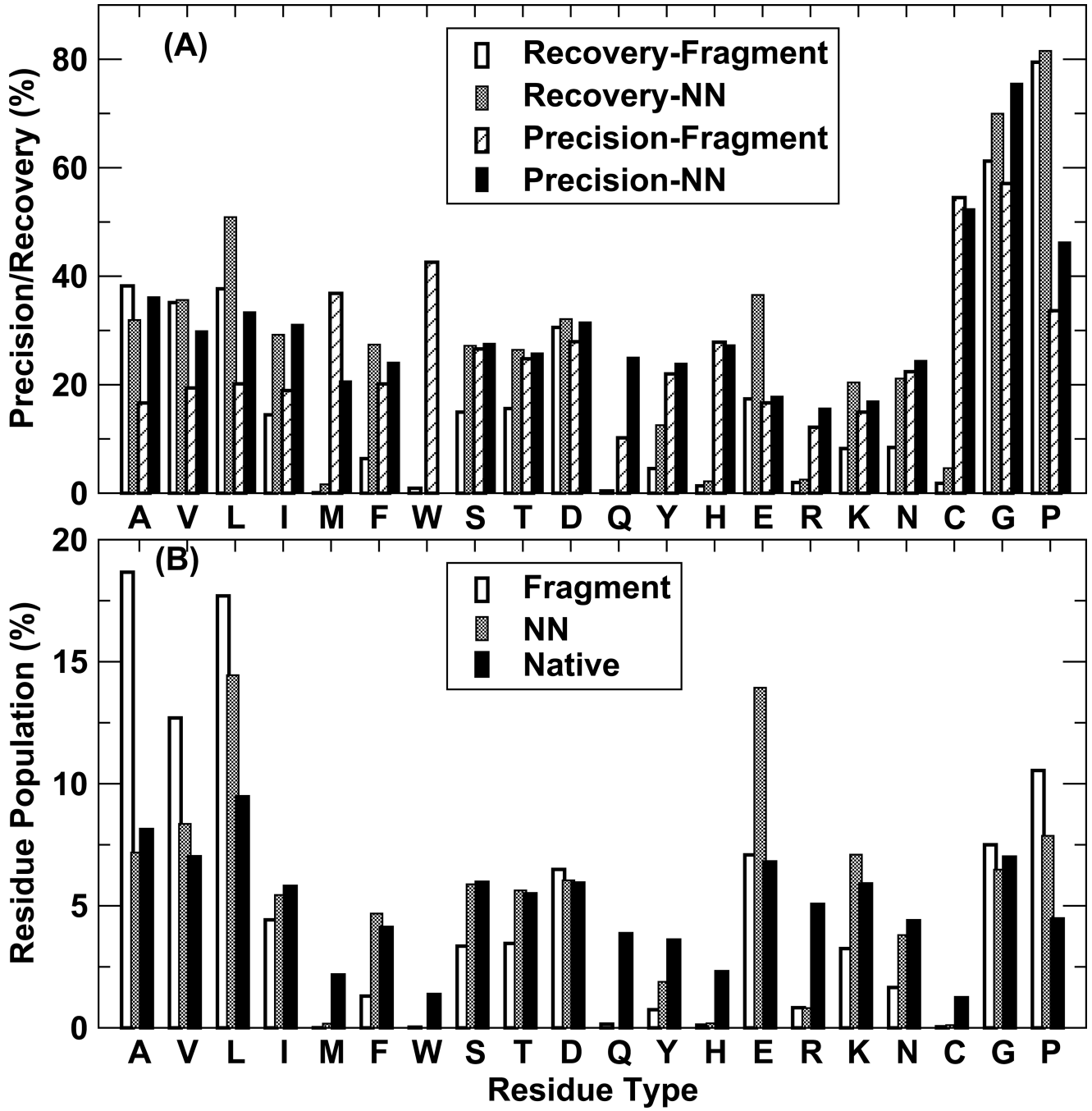

Figure 3.

(A) Recovery rate and precision for each amino acid residue type by fragment-based and neural-network-based approaches as labeled. (B) Frequencies of 20 types of amino acid residues by fragment-based and NN-based approaches are compared to those from wild-type sequences as labeled.

Table I

Sequence identities between predicted and wild-type sequences along with the fraction of hydrophilic residues (the number in parentheses) in different secondary structure, surface (residues with 20% or more solvent accessible surface) and core regions for the independent test set

%	H[a] (f_h[b])	S[a] (f_h)	C[a] (f_h)	Surf (f_h)	Core (f_h)	Lc[c]
Fragment-Based	18.1 (24.9)	19.6 (16.8)	29.9 (37.4)	22.1 (33.5)	26.2 (17.8)	50.8
Neural-Network	26.1 (43.7)	24.5 (30.2)	35.0 (47.7)	26.2 (60.0)	36.7 (17.8)	34.5
Wild-Type	100 (52.2)	100 (35.7)	100 (53.6)	100 (64.7)	100 (27.6)	3

[a] H, S, and C denote helix, sheet and coil, respectively.

[b] f_h denotes fraction of hydrophilic residues (D, E, H, K, N, Q, R, S, T, and Y).

[c] fraction of residues in low complexity regions.

Table II

Performance of various methods measured according to sequence identity to wild-type sequences, consensus sequences from PSSM (either top 1 match or either of the top 2 match) and mean-square error (MSE) to PSSM on the dataset of TR1532 or TS500 (the number in parentheses).

Method	Top 1 (%) TR1532(TS500)	Top 2 (%) TR1532(TS500)	MSE TR1532(TS500)
Fragment-Based	21.5 (21.5)	42.7 (41.8)	0.24 (0.24)
NN (Single)	26.6 (26.3)	51.7 (51.7)	0.21 (0.21)
NN (PSSM)	26.1 (26.7)	50.3 (50.7)	0.18 (0.18)

Table III

Comparison of predicted sequence profiles with wild type sequence or profile for a dataset of randomly selected 50 small proteins with sequence length between 60–200 and fraction of surface residue between 0.5–0.8.

	MSE[a]	Seqid(C,S)[b]	%lc[c]	F_h(C,S)[d]
Fragment-based	0.230	23.4 (24.0, 20.6)	50.4	15.8, 34.6
RosettaDesign	0.223	30.0[e] (45.2, 23.1)	7.1	33.7, 65.2
NN (Single)	0.198	30.3 (37.6, 25.5)	28.5	18.7, 58.4
NN (PSSM)	0.177	27.3 (33.1 ,23.4)	36.1	16.9, 64.5
Wild-Type	0	100 (100, 100)	3.7	26.5, 66.2

[a] The mean square error between predicted and actual PSSM.

[b] The average sequence identity between predicted consensus sequence and wild-type sequence for NN methods. The seqid for RosettaDesign is based on the average seqid of 1000 designed sequences. The numbers in parentheses are sequence identities for core and surface regions of proteins, respectively.

[c] The average fraction of low complexity residues per protein. For RosettaDesign it is based on consensus sequence of 1000 designed sequences.

[d] The fraction of predicted hydrophilic residues in consensus sequences in core and surface of proteins, respectively.

[e] The average sequence identity from 1000 designed sequences.